Sistema de Recomendación Basado en Contenido para el Catálogo de Netflix

Jhon Henry Quintero, Wilfer Esteban García, Pascual Gómez

***Abstract*—Para combatir la sobrecarga de opciones en catálogos de streaming como Netflix, este proyecto desarrolla un sistema de recomendación basado en contenido. El modelo utiliza la vectorización TF-IDF para convertir las características textuales de cada título (descripción, género, elenco) en un formato numérico y calcula la similitud del coseno para encontrar contenido afín. El sistema implementado genera con éxito listas de recomendaciones coherentes, demostrando ser una solución efectiva para mejorar el descubrimiento de contenido y el engagement del usuario.**

***Index Terms*—Sistema de recomendación, filtrado basado en contenido, TF-IDF, similitud del coseno, Netflix, Procesamiento de Lenguaje Natural (PLN).**

I. Introducción

En la era del streaming, la vasta cantidad de contenido en plataformas como Netflix genera la ”paradoja de la elección”, dificultando el descubrimiento de títulos relevantes para el usuario. Para abordar este desafío, el presente trabajo detalla el diseño y la implementación de un sistema de recomendación basado en contenido, utilizando técnicas de Procesamiento de Lengajue Natural. Se modelan las características textuales de cada título—como la descripción, el género y el elenco—transformándolas en vectores numéricos mediante el algoritmo TF-IDF, y la afinidad entre títulos se determina calculando la similitud del coseno entre dichos vectores. El objetivo es demostrar la eficacia de este enfoque para generar recomendaciones coherentes y personalizadas, mejorando así la experiencia del usuario y su engagement con la plataforma.

II. Problema

A. Definición

En la era del streaming, las plataformas Over-the-Top (OTT) como Netflix han acumulado catálogos con miles de títulos. Si bien esta vasta selección es un atractivo clave, también presenta un desafío significativo para el usuario: la sobrecarga de información, también conocida como la ”paradoja de la elección” [1]. Los usuarios a menudo invierten un tiempo considerable navegando sin rumbo, lo que puede llevar a la fatiga de decisión y, en última instancia, a una menor satisfacción y retención. El problema central que este proyecto aborda es cómo mitigar esta sobrecarga mediante la creación de un sistema de recomendación que pueda guiar a los usuarios hacia contenido relevante de manera eficiente, basándose únicamente en las características intrínsecas de los títulos disponibles.

III. Trabajos Relacionados

A continuación, se detallan trabajos previos que abordan problemas similares y las metodologías fundamentales que inspiran este proyecto.

A. Ejemplos con Problemas Similares

El desafío de guiar a los usuarios a través de grandes catálogos no es exclusivo de Netflix. Históricamente, plataformas de recomendación de películas como MovieLens han sido un campo de pruebas para estos algoritmos. Un análisis de Lam et al. sobre este dataset exploró cómo diferentes enfoques pueden solucionar el problema del ”arranque en frío” (cold start), que ocurre cuando un ítem nuevo (o un usuario nuevo) no tiene un historial de interacciones [2]. Su trabajo es relevante porque nuestro sistema, al no usar datos de usuarios, trata cada recomendación como un problema de cold start, basándose únicamente en los metadatos del contenido. De manera similar, gigantes del e-commerce como Amazon utilizan técnicas de recomendación basadas en atributos de productos para facilitar el descubrimiento, validando la eficacia de analizar las características del ítem para mejorar la experiencia del usuario.

B. Metodologías para la Solución

La base teórica de nuestro sistema se fundamenta en metodologías de recuperación de información y procesamiento de lenguaje natural (PLN).

- **Modelo de Espacio Vectorial y TF-IDF:** El concepto de representar documentos como vectores numéricos fue formalizado por Salton et al. [3]. Este es el pilar de nuestro enfoque, donde cada título de Netflix se trata como un ”documento”. Para ponderar la importancia de las palabras en la descripción, género, etc., se utiliza el esquema TF-IDF (Term Frequency-Inverse Document Frequency). Esta técnica ha demostrado ser un estándar en la industria para la clasificación de texto y la recuperación de información.
- **Modelado Basado en Características y Similitud del Coseno:** Un estudio clave de Pazzani y Billsus demostró cómo se pueden aprender perfiles de ítems modelando sus características intrínsecas para luego inferir similitudes [4]. Nuestro proyecto adopta este principio directamente: creamos un ”perfil de contenido” para cada título y luego medimos qué tan parecidos son. Para cuantificar esta similitud, empleamos la similitud del coseno, una métrica robusta que calcula el coseno del ángulo entre dos vectores en un espacio multidimensional.

IV. Datos y Metodología

A. Objetivos del Análisis y Métricas de Éxito

El Análisis Exploratorio de Datos (EDA) se guía por el problema de negocio central: la ”paradoja de la elección” en Netflix. Con un catálogo tan extenso, los usuarios sufren de fatiga de decisión, lo que aumenta el riesgo de abandono de la plataforma (churn). Para mitigar esto, el proyecto propone un sistema de recomendación basado en contenido que sugiere títulos similares basándose en sus características intrínsecas (género, descripción, elenco), un enfoque de Machine Learning necesario para capturar los matices que las reglas simples no pueden.

El éxito del modelo se medirá con dos métricas clave:

- **Métrica de Negocio:** Aumentar el engagement del usuario, buscando que al menos el 20% del contenido visto provenga de una recomendación directa.
- **Métrica Técnica:** Alcanzar un 85% en la Tasa de Acierto de Relevancia en el Top-5 (Relevance Hit Rate @5). Un ”acierto” se define como una recomendación donde al menos 2 de los 5 títulos sugeridos comparten un subgénero, actor o director clave con el título original.

B. Descripción del Conjunto de Datos

Para este proyecto se utilizó el dataset público ”Netflix Movies and TV Shows”, originado por Flixable y disponible en Kaggle, que contiene aproximadamente 8,800 títulos del catálogo de Netflix hasta 2021. Cada fila representa un título único (película o serie). Este conjunto de datos fue seleccionado por su tamaño manejable para el prototipado, su riqueza de tipos de datos (numéricos, categóricos y textuales) y porque presenta desafíos de limpieza realistas, como la presencia de valores nulos en columnas clave como director y cast. Estas características lo convierten en un entorno ideal para desarrollar y evaluar un sistema de recomendación basado en contenido. El archivo de datos se gestiona localmente en una carpeta datasets/ excluida del control de versiones.

V. Análisis Exploratorio de Datos

A. Estructura Inicial y Diccionario de Datos

El primer paso del análisis consiste en una inspección de la estructura general del dataset. El conjunto de datos se compone de 8,807 registros y 12 columnas. La Tabla 1 resume las características de cada variable, su tipo de dato y la cantidad de valores no nulos, lo que inmediatamente destaca los desafíos de preprocesamiento que se abordarán más adelante.

De la tabla 1. se extraen dos conclusiones fundamentales para el proyecto:

- **Presencia de Valores Nulos:** Las columnas cruciales para la recomendación, director y cast, presentan una cantidad significativa de datos faltantes, estableciendo el manejo de estos nulos como una tarea prioritaria.
- **Riqueza Textual:** Las columnas listed_in (géneros) y description (sinopsis) están completas y contienen la información semántica que será el insumo principal para el modelo de recomendación.

Variable	Tipo de Dato	No Nulos	Descripción
show_id	object	8807	Identificador único del título.
type	object	8807	Indica si es ’Movie’ o ’TV Show’.
title	object	8807	Nombre del título.
director	object	6173	Director(es) de la obra.
cast	object	7982	Elenco principal.
country	object	7976	País de producción.
date_added	object	8797	Fecha de adición a Netflix.
release_year	int64	8807	Año de estreno original.
rating	object	8803	Clasificación por edad.
duration	object	8804	Duración o número de temporadas.
listed_in	object	8807	Géneros.
description	object	8807	Sinopsis.

TABLE I
Diccionario de Datos del Catálogo de Netflix

B. Descripción de Variables Clave

Las variables del dataset se pueden agrupar en categorías funcionales para guiar el análisis. Las variables de contenido son el núcleo del modelo de recomendación: description, listed_in (géneros), director y cast proveen la riqueza semántica para calcular la similitud. Por otro lado, las variables de clasificación como type (Movie/TV Show), rating (clasificación por edad) y duration permiten segmentar y filtrar el catálogo. Finalmente, las variables temporales (release_year, date_added) y geográficas (country) ofrecen un contexto adicional sobre la producción y distribución del contenido, posibilitando el análisis de tendencias históricas y culturales.

C. Análisis de Valores Faltantes

Un paso crítico en el EDA es la cuantificación de datos ausentes, ya que pueden degradar el rendimiento del modelo. El análisis revela que tres columnas clave para la recomendación presentan un volumen considerable de valores faltantes (ver Figura 1).

La columna director es la más afectada, con casi un 30% de sus datos ausentes. Le siguen country y cast, ambas con aproximadamente un 9% de valores nulos. Esta ausencia de datos es un problema a resolver, dado que estas características son insumos importantes para la creación del ”perfil de contenido”.

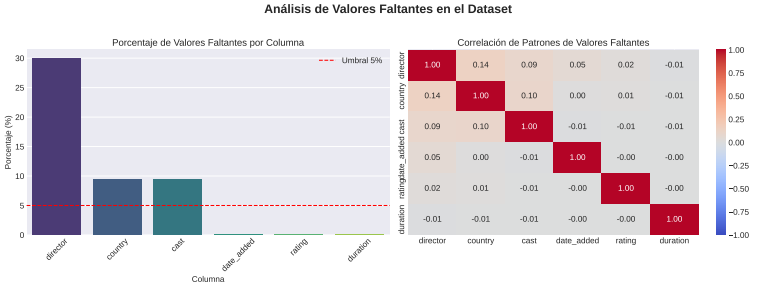


Fig. 1. Análisis de valores faltantes. (Izquierda) Porcentaje de datos ausentes por cada columna afectada. (Derecha) Matriz de correlación que muestra la tendencia de los valores a estar ausentes conjuntamente.

La matriz de correlación de patrones faltantes (Fig. 1, derecha) muestra una correlación positiva débil (0.14) entre

la ausencia de director y country, sugiriendo que cuando falta el director, hay una ligera tendencia a que también falte el país de producción.

Debido a que eliminar casi un tercio del dataset (las filas donde falta el director) no es viable y estas columnas son importantes, la estrategia a seguir será la imputación: se rellenarán los valores nulos con un marcador textual (ej. 'Desconocido') para preservar los registros y asegurar que el modelo de PLN pueda procesarlos.

D. Análisis de Distribuciones Numéricas

El análisis univariado de las variables numéricas release_year y duration_min revela características importantes sobre la composición del catálogo de películas de Netflix.

1) **Distribución del Año de Estreno (release_year):** El histograma del año de estreno (ver Figura 2) muestra una distribución con una fuerte asimetría negativa (sesgo a la izquierda). La media (2014.18) es considerablemente menor que la mediana (2017.00), lo que indica que la gran mayoría de los títulos son muy recientes, con una larga cola de películas más antiguas que actúan como outliers. Esto confirma la estrategia de Netflix de enfocarse en contenido moderno, manteniendo un selecto catálogo de clásicos.

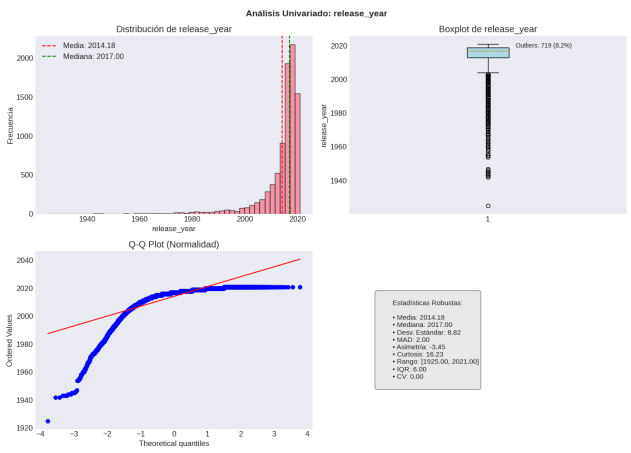


Fig. 2. Análisis univariado de la variable release_year, mostrando la distribución, boxplot y normalidad.

2) **Distribución de la Duración de las Películas (duration_min):** La duración de las películas (ver Figura 3) presenta una distribución que se asemeja más a una curva normal, aunque con un ligero sesgo positivo (a la derecha). La media (99.58 min) y la mediana (98.00 min) son muy cercanas, lo que sugiere que la mayoría de las películas se agrupan en torno a la duración estándar de 90-100 minutos. Los outliers detectados corresponden a cortometrajes (¡ 40 min) y películas épicas (¿ 180 min), representando la diversidad de formatos en la plataforma.

El análisis de correlación entre ambas variables muestra un coeficiente de Pearson de -0.206, lo que indica una correlación negativa muy débil. En la práctica, esto significa que no existe una relación significativa entre el año de estreno de una película y su duración en el catálogo de Netflix.

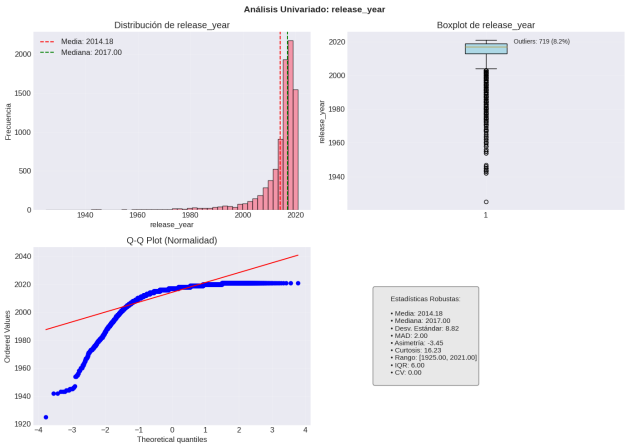


Fig. 3. Análisis univariado de la variable duration_min.

E. Análisis de Variables Categóricas Clave

El análisis de las variables categóricas principales, type y rating, revela la estrategia de contenido de Netflix y destaca la necesidad de limpieza de datos.

1) **Distribución por Tipo de Contenido (type):** El catálogo de Netflix muestra una clara dominancia de películas ('Movie'), que constituyen casi el 70% del total, frente a un 30% de series de televisión ('TV Show'). Un hallazgo clave (ver Figura 4) es que la mediana del año de estreno de las series es notablemente más reciente que la de las películas. Esto sugiere que, si bien Netflix tiene un catálogo de películas más amplio y con mayor recorrido histórico, su estrategia de producción de contenido original se ha centrado fuertemente en el formato de series en los últimos años.

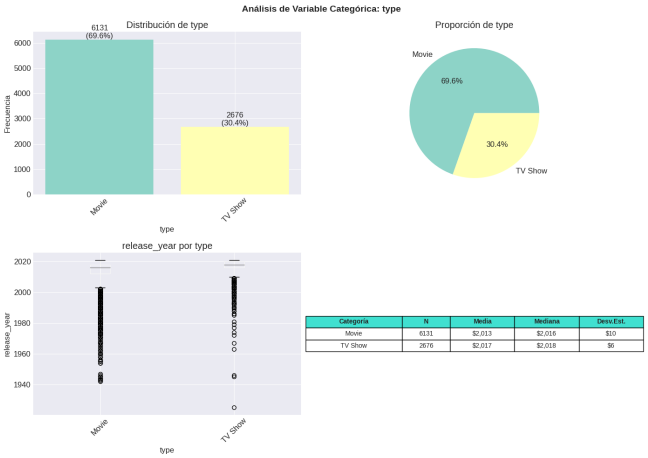
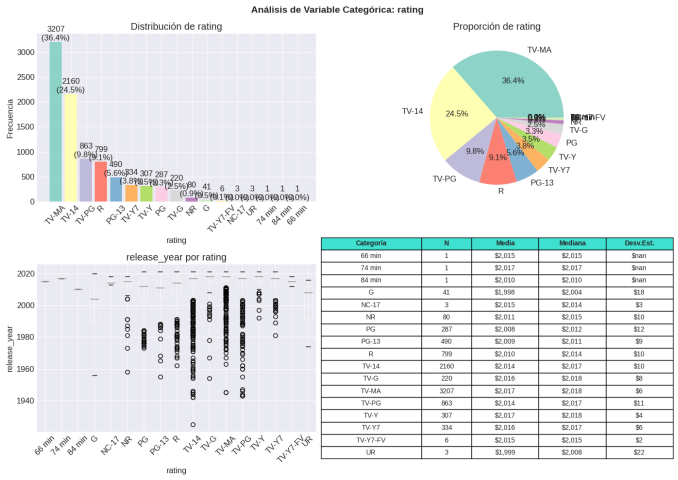


Fig. 4. Análisis de la variable type, mostrando la distribución y la relación con el año de estreno.

2) **Distribución por Clasificación de Edad (rating):** La clasificación por edad está concentrada en las categorías para audiencias maduras y adolescentes. TV-MA (Mature Audience) es la categoría más frecuente con un 36.4%, seguida por TV-14 (Parents Strongly Cautioned) con un 24.5%. Esto indica que el contenido principal de la plataforma está dirigido a un público adulto y juvenil. Además, el análisis (ver Figura 5) expone la presencia de datos ruidosos en esta columna, como valores que corresponden a duraciones de películas (”66 min”, ”74 min”), lo que requerirá un paso de limpieza y estandarización antes del modelado.



Categoría	N	Media	Mediana	Desv.Est.
66 min	1	$2,015	$2,015	$nan
74 min	1	$2,017	$2,017	$nan
84 min	1	$2,010	$2,010	$nan
G	41	$1,998	$2,004	$18
NC-17	3	$2,015	$2,014	$3
NR	80	$2,011	$2,015	$10
PG	287	$2,008	$2,012	$12
PG-13	490	$2,009	$2,011	$9
R	799	$2,010	$2,014	$10
TV-14	2160	$2,014	$2,017	$10
TV-G	220	$2,016	$2,018	$8
TV-MA	3207	$2,017	$2,018	$6
TV-PG	863	$2,014	$2,017	$11
TV-Y	307	$2,017	$2,018	$4
TV-Y7	334	$2,016	$2,017	$6
TV-Y7-FV	6	$2,015	$2,015	$2
UR	3	$1,999	$2,008	$22

Fig. 5. Análisis de la variable rating, mostrando una gran cantidad de categorías y datos ruidosos.

F. Análisis Geográfico de la Producción de Contenido

El análisis de la variable country permite entender el origen geográfico del contenido y revela la estrategia de mercado global de Netflix.

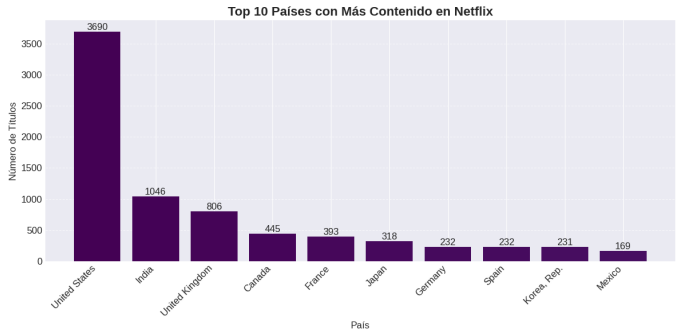


Fig. 6. Top 10 de países con mayor número de títulos en el catálogo de Netflix.

El análisis muestra una fuerte concentración de la producción en Estados Unidos, que domina el catálogo con más de 3,600 títulos, triplicando al segundo país, India (ver Figura 6). El Top 10 está compuesto principalmente por países de América del Norte, Europa y Asia, destacando potencias cinematográficas como el Reino Unido, Japón y Corea del Sur.

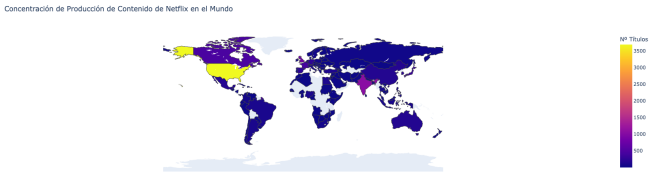


Fig. 7. Mapa de concentración de la producción de contenido a nivel mundial.

El mapa mundial (Figura 7) refuerza esta visión, ilustrando cómo la producción de contenido no está distribuida de manera uniforme, sino concentrada en hubs de producción específicos. Esto subraya la importancia de la variable country como una característica potencialmente útil para la recomendación, permitiendo agrupar contenido por origen cultural y estilístico (ej. ”cine de Hollywood”, ”anime japonés” o ”Bollywood”).

G. Análisis de Correlación entre Variables Numéricas

Para finalizar el EDA, se analiza la relación entre las dos únicas variables numéricas del estudio: release_year y duration_min. Este paso busca determinar si existe alguna dependencia entre la antigüedad de una película y su duración.

El análisis de correlación de Pearson, que mide relaciones lineales, arroja un coeficiente de -0.21. Por su parte, la correlación de Spearman, que mide relaciones monotónicas (si una variable tiende a aumentar o disminuir cuando la otra lo hace, sin ser necesariamente lineal), da un valor de -0.19 (ver Figura 8).

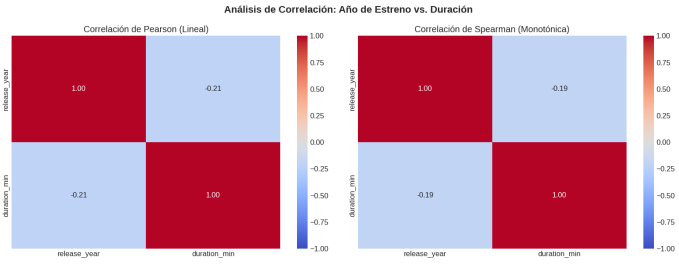


Fig. 8. Mapas de calor de correlación de Pearson (izquierda) y Spearman (derecha) entre el año de estreno y la duración de las películas.

H. Detección de Anomalías y Outliers

El análisis final del EDA se enfoca en la detección de valores atípicos (outliers) en las variables numéricas para determinar si representan errores de datos o información valiosa. Se utilizaron tres métodos: Rango Intercuartílico (IQR), Z-Score e Isolation Forest.

Los resultados (ver Figura 9) muestran que ambas variables, release_year y duration_min, contienen una cantidad considerable de outliers. El método IQR, por ejemplo, identifica más de 500 outliers en release_year y 450 en duration_min.

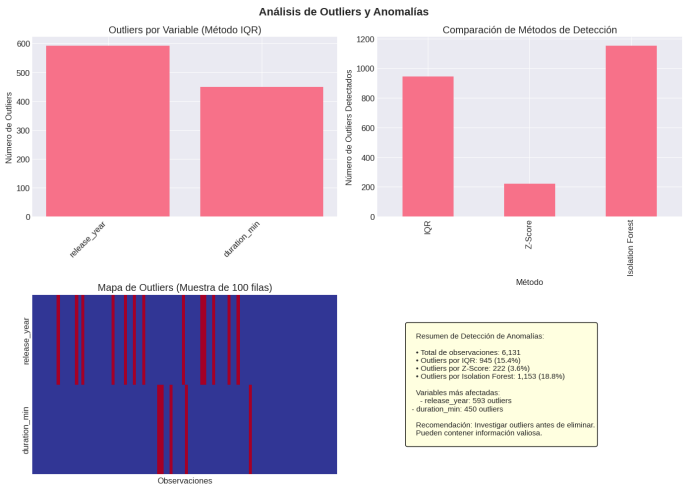


Fig. 9. Dashboard de detección de anomalías en las variables numéricas del catálogo de películas.

VI. Baseline y Conclusiones Preliminares

A. *Baseline del Modelo*

Antes de construir el modelo de Machine Learning, se define un modelo base (baseline) simple para tener un punto de comparación. Para este problema, un baseline razonable sería: ”Dado un título, recomendar las 5 películas o series más populares que compartan su género principal”. Este enfoque no requiere un análisis semántico profundo y sirve como un estándar mínimo a superar. Nuestro modelo de TF-IDF deberá ofrecer recomendaciones más específicas y de nicho que esta simple regla.

B. *Conclusiones del Análisis Exploratorio de Datos*

El análisis exploratorio del catálogo de Netflix ha revelado varios puntos clave que informarán el preprocesamiento de datos y la construcción del modelo:

- **El Catálogo está Dominado por Contenido Moderno:** La distribución de release_year está fuertemente sesgada, confirmando que la mayoría del contenido es reciente. Las películas clásicas existen, pero son consideradas outliers.
- **Foco en Audiencias Maduras y Juveniles:** Las clasificaciones de edad más comunes son TV-MA y TV-14, lo que indica una estrategia de contenido dirigida principalmente a adultos y adolescentes.
- **Se Requiere Limpieza de Datos Significativa:** El análisis destacó la necesidad de manejar una cantidad considerable de valores nulos (especialmente en director y cast), datos ruidosos (en la columna rating) y la estandarización de datos categóricos (como country).
- **No Hay Correlación entre Antigüedad y Duración:** Se demostró que no existe una relación significativa entre el año de estreno de una película y su duración.

En resumen, el EDA confirma que el dataset es idóneo para un sistema de recomendación basado en contenido gracias a su riqueza textual, pero subraya que un preprocesamiento de datos robusto es un paso crítico e indispensable para el éxito del modelo.

References

[1] B. Schwartz, The Paradox of Choice: Why More Is Less. New York: Ecco, 2004.
[2] X. Lam, T. Vu, T. D. Bui, and D. Phung, ”Addressing the Cold-Start Problem in Recommendation Systems,” in Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, Suwon, Korea, 2008, pp. 209-213.
[3] G. Salton, A. Wong, and C. S. Y. Yang, ”A Vector Space Model for Automatic Indexing,” Communications of the ACM, vol. 18, no. 11, pp. 613–620, 1975.
[4] M. J. Pazzani and D. Billsus, ”Learning and Revising User Profiles: The Identification of Interesting Web Sites,” Machine Learning, vol. 27, pp. 313-331, 1997.

VII. Repositorio

https://github.com/mr-sudaca/curso_ml